

DOUGLAS A. WINTHROP (Bar No. 183532)
Douglas.Winthrop@arnoldporter.com
JOSEPH FARRIS (Bar No. 263405)
Joseph.Farris@arnoldporter.com
JESSICA L. GILLOTTE (Bar No. 333517)
Jessica.Gillotte@arnoldporter.com
ESTAYVAINE BRAGG (Bar No. 341400)
Estayvaine.Bragg@arnoldporter.com
ARNOLD & PORTER KAYE SCHOLER LLP
Three Embarcadero Center, 10th Floor
San Francisco, CA 94111-4024
Telephone: (415) 471-3100
Facsimile: (415) 471-3400

MARK LEMLEY (Bar No. 155830)
mlemley@lex-lumina.com
LEX LUMINA LLP
700 S. Flower Street, Suite 1000
Los Angeles, CA 90017
Telephone: (213) 600-6063

Attorneys for Defendant ANTHROPIC PBC

ANGEL T. NAKAMURA (Bar No. 205396)
Angel.Nakamura@arnoldporter.com
OSCAR RAMALLO (Bar No. 241487)
Oscar.Ramallo@arnoldporter.com
ALLYSON MYERS (Bar No. 342038)
Ally.Myers@arnoldporter.com
ARNOLD & PORTER KAYE SCHOLER LLP
777 South Figueroa Street, 44th Floor
Los Angeles, CA 90017-5844
Telephone: (213) 243-4000
Facsimile: (213) 243-4199

JOSEPH R. WETZEL (Bar No. 238008)
joe.wetzel@lw.com
ANDREW M. GASS (Bar No. 259694)
andrew.gass@lw.com
LATHAM & WATKINS LLP
505 Montgomery Street, Suite 2000
San Francisco, CA 94111
Telephone: (415) 391-0600
Facsimile: (415) 395-8095

UNITED STATES DISTRICT COURT
NORTHERN DISTRICT OF CALIFORNIA
SAN FRANCISCO DIVISION

ANDREA BARTZ, ANDREA BARTZ, INC.,
CHARLES GRAEBER, KIRK WALLACE
JOHNSON, and MJ + KJ, INC., individually and on
behalf of others similarly situated,

Plaintiffs,

v.

ANTHROPIC PBC,

Defendant.

Case No. 3:24-CV-05417-WHA

Action Filed: August 19, 2024

**DEFENDANT ANTHROPIC PBC'S
NOTICE OF MOTION AND MOTION
FOR SUMMARY JUDGMENT;
MEMORANDUM OF POINTS AND
AUTHORITIES IN SUPPORT**

Date: May 15, 2025

Time: 8:00 a.m.

Place: Courtroom 12, 19th Floor

Judge: Honorable William H. Alsup

TABLE OF CONTENTS

	<u>Page</u>
NOTICE OF MOTION AND MOTION	viii
MEMORANDUM OF POINTS AND AUTHORITIES	1
STATEMENT OF ISSUES TO BE DECIDED	1
INTRODUCTION AND SUMMARY OF ARGUMENT	1
SUMMARY OF UNDISPUTED MATERIAL FACTS.....	2
A. Anthropic and its mission.	2
B. How people use Claude.....	3
C. LLM training.....	3
D. This case is only about use of Plaintiffs’ books as training data.	7
E. Anthropic’s use of books as training data.	7
F. Licensing is not a practicable approach to assembling the datasets, or even just the books datasets, required to build Claude.....	8
LEGAL STANDARD	10
ARGUMENT	10
I. THE FIRST FACTOR FAVORS FAIR USE BECAUSE USING COPYRIGHTED WORKS TO TRAIN AN LLM IS QUINTESENTIALLY TRANSFORMATIVE	11
II. THE SECOND FACTOR FAVORS FAIR USE BECAUSE ANTHROPIC USES PLAINTIFFS’ PUBLISHED WORKS FOR A TRANSFORMATIVE PURPOSE	17
III. THE THIRD FACTOR FAVORS FAIR USE BECAUSE ANTHROPIC’S TRANSFORMATIVE PURPOSE REQUIRES COPYING ENTIRE WORKS.....	17
IV. THE FOURTH FACTOR FAVORS FAIR USE BECAUSE ANTHROPIC’S TRANSFORMATIVE COPYING DOES NOT HARM ANY COGNIZABLE MARKET	18
A. There is no cognizable market harm from Anthropic’s transformative use as a matter of law.....	18
B. There is no viable existing or potential market to license Plaintiffs’ works for LLM training.	20
C. The public benefits of Claude outweigh any potential market effect.	23
CONCLUSION	24

TABLE OF AUTHORITIES

Page(s)

Cases

<i>A.V. ex rel. Vanderhye v. iParadigms, LLC</i> , 562 F.3d 630 (4th Cir. 2009).....	11, 14, 18
<i>Accord Concord Music Grp., Inc., et al. v. Anthropic PBC</i> , No. 24cv-03811-EKL (N.D. Cal. March 25, 2025)	14
<i>Anderson v. Liberty Lobby, Inc.</i> , 477 U.S. 242 1986)	10
<i>Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith</i> , 598 U.S. 508 (2023)	12, 13
<i>Apple Inc. v. Corellium, Inc.</i> , 2023 WL 3295671 (11th Cir. May 8, 2023)	15
<i>Authors Guild v. Google, Inc.</i> , 804 F.3d 202 (2d Cir. 2015).....	<i>passim</i>
<i>Authors Guild, Inc. v. HathiTrust</i> , 755 F.3d 87, 97 (2d Cir. 2014).....	12, 13, 17, 19
<i>Bell v. Eagle Mountain Saginaw Indep. Sch. Dist.</i> , 27 F.4th 313 (5th Cir. 2022)	20
<i>Bell v. Milwaukee Bd. of Sch. Dirs.</i> , 2022 WL 18276966 (E.D. Wis. Dec. 21, 2022).....	21
<i>Bill Graham Archives v. Dorling Kindersley Ltd.</i> , 448 F.3d 605 (2d Cir. 2006).....	17, 19
<i>Cambridge Univ. Press v. Patton</i> , 769 F.3d 1232 (11th Cir. 2014).....	20
<i>Campbell v. Acuff-Rose Music, Inc.</i> , 510 U.S. 569 (1994).....	<i>passim</i>
<i>Campbell. Oracle Am., Inc. v. Google Inc.</i> , 2016 WL 3181206 (N.D. Cal. June 8, 2016)	16
<i>Google LLC v. Oracle Am., Inc.</i> , 593 U.S. 1 (2021).....	<i>passim</i>
<i>Harper & Row Publishers, Inc. v. Nation Enterprises</i> , 471 U.S. 5393 (1985).....	16

1	<i>Kelly v. Arriba Soft Corp.</i> ,	
2	336 F.3d 811 (9th Cir. 2003).....	11, 15, 17
3	<i>Mattel, Inc. v. Walking Mountain Prods.</i> ,	
4	353 F.3d 792 (9th Cir. 2003).....	17
5	<i>Perfect 10, Inc. v. Amazon.com, Inc.</i> ,	
6	508 F.3d 1146 (9th Cir. 2007).....	<i>passim</i>
7	<i>Sega Enters. Ltd. v. Accolade, Inc.</i> ,	
8	977 F.2d 1510 (9th Cir. 1992).....	<i>passim</i>
9	<i>Seltzer v. Green Day, Inc.</i> ,	
10	725 F.3d 1170 (9th Cir. 2013).....	19, 20, 21, 23
11	<i>SOFA Ent., Inc. v. Dodger Prods., Inc.</i> ,	
12	709 F.3d 1273 (9th Cir. 2013).....	23
13	<i>Sony Comput. Ent., Inc. v. Connectix Corp.</i> ,	
14	203 F.3d 596 (9th Cir. 2000).....	11, 14, 18, 20
15	<i>Thomson Reuters Enter. Ctr. GMBH v. ROSS Intelligence Inc.</i> ,	
16	694 F. Supp. 3d 467 (D. Del. 2023).....	14
17	<i>Thomson Reuters Enterprise Center GMBH v. ROSS Intelligence Inc.</i> ,	
18	2025 WL 458520 (D. Del. Feb. 11, 2025).....	14
19	<i>Time Inc. v. Bernard Geis Associates</i> ,	
20	293 F. Supp. 130 (S.D.N.Y. 1968).....	16
21	<i>Tresóna Multimedia, LLC v. Burbank High Sch. Vocal Music Ass’n</i> ,	
22	953 F.3d 638 (9th Cir. 2020).....	19
23	<i>Ty, Inc. v. Publ’ns Int’l Ltd.</i> ,	
24	292 F.3d 512 (7th Cir. 2002).....	21
25	<i>White v. West Pub. Corp.</i> ,	
26	29 F. Supp. 3d 396 (S.D.N.Y. 2014).....	18, 21, 23

Other Authorities


27	17 U.S.C. § 107	1, 10, 17, 18
28	Pierre N. Leval, <i>Toward A Fair Use Standard</i> , 103 Harv. L. Rev. 1105, 1111 (1990)...	11, 16, 19
	U.S. Const. art. 1, § 8, cl. 8	20

TABLE OF EXHIBITS

Exhibit	Description
Decl.	Declaration of Jared Kaplan
Decl.	Declaration of Tom Turvey
Decl.	Declaration of Steven Peterson
1	March 5, 2025 Deposition Transcript of Charles Andrew Graeber
2	March 7, 2025 Deposition Transcript of Andrea Marie Bartz
3	March 6, 2025 Deposition Transcript of Kirk Wallace Johnson
4	Plaintiff Charles Graeber's January 28, 2025 Responses to Defendant's Request for Admission No. 9
5	Plaintiffs Andrea Bartz and Andrea Bartz, Inc.'s January 28, 2025 Responses to Defendant's Request for Admission No. 9
6	Plaintiffs Kirk Wallace Johnson and MJ + KJ, Inc.'s January 28, 2025 Responses to Defendant's Request for Admission No. 9
7	A document entitled "Concise Research Roadmap" (ANT_BARTZ_000436674)
8	Trenton Bricken et al., <i>Towards Monosemanticity: Decomposing Language Models With Dictionary Learning</i> (Oct. 4, 2023)
9	Adly Templeton et al., <i>Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet</i> (May 21, 2024)
10	Arc Institute, <i>Evo Mechanistic Interpretability Visualizer</i>
11	Garyk Brixi et al., <i>Genome Modeling and Design Across All Domains of Life with Evo 2</i> (Feb. 21, 2025)
12	Elana Simon & James Zou, <i>InterPLM: Discovering Interpretable Features in Protein Language Models via Sparse Autoencoders</i> (Jan. 28, 2025)
13	Yuntao Bai et al., <i>Constitutional AI: Harmlessness from AI Feedback</i> (Dec. 15, 2022)
14	Google Scholar Citations for Yuntao Bai et al., <i>Constitutional AI: Harmlessness from AI Feedback</i> (Dec. 15, 2022)

1	15	Cem Anil et al., <i>Many-shot Jailbreaking</i> (Apr. 2, 2024) (ANT_BARTZ_000000544-577)
2	16	Alex Tamkin et al., <i>Evaluating and Mitigating Discrimination in Language Model Decisions</i> (Dec. 6, 2023)
3	17	Erik Jones et al., <i>Forecasting Rare Language Model Behaviors</i> (Feb. 24, 2025)
4	18	Alex Tamkin et al., <i>Clio: Privacy-Preserving Insights into Real World AI Use</i> (Dec. 18, 2024)
5	19	Anthropic, <i>Claude 3.5 Sonnet for Sparking Creativity</i> (Jun. 20, 2024)
6	20	Anthropic, <i>Introducing Claude</i> (Mar. 14, 2023)
7	21	Anthropic, <i>Claude 3.5 Sonnet as a Writing Partner</i> (June 20, 2024) (ANT_BARTZ_000436709)
8	22	Anthropic, <i>The University of Sydney and Accenture Accelerate Whale Conservation with Claude</i>
9	23	Anthropic, <i>MagicSchool Transforms K-12 Education for 3 Million Educators and Their Students with Claude</i>
10	24	Anthropic, <i>European Parliament Expands Access to Their Archives with Claude</i>
11	25	Anthropic, <i>Anthropic & AWS: Enterprise AI at scale</i>
12	26	Anthropic, <i>Steno Helps Attorneys Find the Critical Insights in Legal Transcripts with Claude</i> (ANT_BARTZ_000002487-492)
13	27	Anthropic, <i>Jumpcut Helps Hollywood Find the Next Big Script with Claude</i> (ANT_BARTZ_0000002385-390)
14	28	Anthropic, <i>Campfire Accelerates Accounting with Claude</i>
15	29	Accenture, AWS, Accenture and Anthropic Join Forces to Help Organizations Scale AI Responsibly (Mar. 20, 2024)
16	30	Danny Hernandez et al., <i>Scaling Laws and Interpretability of Learning from Repeated Data</i> (May 21, 2022)
17	31	Chiyuan Zhang et al., <i>Counterfactual Memorization in Neural Language Models</i> (Sept. 22, 2022)
18	32	Yiheng Liu et al., <i>Understanding LLMs: A Comprehensive Overview from Training to Inference</i> (Jan. 6, 2024)

-v-



33	Jordan Hoffmann et al., <i>Training Compute-Optimal Large Language Models</i> (Mar. 29, 2022)
34	Brando Miranda et al., <i>Beyond Scale: The Diversity Coefficient as a Data Quality Metric for Variability in Natural Language Data</i> (Aug. 26, 2024)
35	Priya Khandelwal et al., <i>A Guide to Improving Long Context Instruction Following on Open Source Models</i> (Sept. 26, 2024)
36	Tianyu Gao et al., <i>How to Train Long-Context Language Models (Effectively)</i> (Oct. 3, 2024)
37	Anthropic, <i>The Claude 3 Model Family: Opus, Sonnet, Haiku 3</i> (2024) (ANT_BARTZ_000000381-422)
38	Anthropic, <i>Model Card and Evaluations for Claude Models 2</i> (2023) (ANT_BARTZ_000000367-80)
39	Leo Gao et al., <i>The Pile: An 800GB Dataset of Diverse Text for Language Modeling</i> (Dec. 31, 2020)
40	Internet Archive, <i>About the Internet Archive</i>
41	Model Data Domain Breakdown Google Sheet (ANT_BARTZ_000435165)
42	Yuntao Bai et al., <i>Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback</i> (Apr. 12, 2022) (ANT_BARTZ_000004461-534)
43	Amanda Askell et al., <i>A General Language Assistant as a Laboratory for Alignment</i> (Dec. 9, 2021) (ANT_BARTZ_000002865-912)
44	Anthropic, <i>Consumer Terms of Service</i> (Feb. 19, 2025)
45	WTP Valuation Model (ANT_BARTZ_000279410)
46	Jay Peters, <i>HarperCollins Is Asking Authors to License Their Books for AI Training</i> (Nov. 18, 2024)
47	Authors Guild, <i>HarperCollins AI Licensing Deal</i> (Nov. 19, 2024)
48	Elsevier, <i>ScienceDirect: Elsevier's Premier Platform of Peer-reviewed Scholarly Literature</i>
49	Jim Millriott, <i>Wiley Wraps Up Divestiture Program, Looks at AI Opportunities</i> (Sept. 5, 2024)



1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28

50	Oxford Academic, <i>Our Book Publishing Programme</i>
----	---

NOTICE OF MOTION AND MOTION

TO ALL PARTIES AND THEIR COUNSEL OF RECORD:

Please take notice that on May 15, 2025 at 8:00 a.m., or as soon thereafter as the matter may be heard, in the courtroom of the Honorable William H. Alsup, Courtroom 12, 19th Floor, San Francisco Courthouse, 450 Golden Gate Avenue, San Francisco, California, Defendant Anthropic PBC (“Anthropic”) will and hereby does move, pursuant to Rule 56 of the Federal Rules of Civil Procedure, for an order granting summary judgment, on all claims asserted in Plaintiffs’ Andrea Bartz, Andrea Bartz, Inc., Charles Graeber, Kirk Wallace Johnson, and MJ + KJ, Inc.’s (“Plaintiffs”) First Amended Class Action Complaint. ECF No. 70.

The Motion should be granted, and summary judgment entered in Anthropic’s favor because Anthropic’s use of Plaintiffs’ works is a fair use under Section 107 of the Copyright Act.

This Motion is based on this Notice of Motion and Motion, the supporting Memorandum of Points and Authorities, the Declarations of Jared Kaplan, Tom Turvey, Steven Peterson, and Douglas A. Winthrop, and all exhibits in support of those declarations, the complete files and records in this action, and such argument and evidence as may be presented before or at the hearing.

MEMORANDUM OF POINTS AND AUTHORITIES

STATEMENT OF ISSUES TO BE DECIDED

Is Anthropic’s use of Plaintiffs’ books to train its large language models (“LLMs”) a protected fair use under Section 107 of the Copyright Act?

INTRODUCTION AND SUMMARY OF ARGUMENT

Anthropic’s LLM, Claude, represents a revolutionary advance in computing. Claude interprets and responds to a wide range of user queries like an intelligent human, allowing users to engage in general, open-ended interactions in the service of an unlimited set of potential projects. In so doing, Claude demonstrates complex reasoning, problem-solving, and creativity across a broad array of tasks from software coding, to writing projects, to data analysis, and beyond. Claude has been used by biology researchers to study protein sequences, by scientists to analyze whale recordings and enhance conservation efforts, by educators to develop teaching tools, by government agencies to summarize and translate vast archival records, by pharmaceutical companies to accelerate clinical trials for therapeutic drugs, and by everyday working professionals in their jobs and lives. It is a radically transformational tool for creators of many kinds—writers, teachers, scientists, businesspeople and more—which enables new expression and innovation to flourish.

To do this, Claude, like other LLMs, must be trained on a vast amount of data (the most recent version used the equivalent of at least [REDACTED] words), including data drawn from books and other writings, all showing how humans think and use language. Plaintiffs in this case are book authors, and there is no dispute that Anthropic has used Plaintiffs’ books as a miniscule part of the corpus of data to train Claude. But Plaintiffs do not contend that Claude will generate a copy of their books, or even something substantially similar. Rather, Plaintiffs’ claim is that Anthropic’s mere use of Plaintiffs’ books to train its LLMs is copyright infringement.

That contention is wrong as a matter of law. Consistent with a long line of cases affirming the right to engage in similar back-end copying in developing new digital technologies, Anthropic’s use of Plaintiffs’ works to train its LLMs is fair. The use serves a fundamentally different purpose from the books themselves. It is, in the language of fair use, a “transformative” use that the Copyright Act not only allows, but encourages. *Perfect 10, Inc. v. Amazon.com, Inc.*, 508 F.3d 1146, 1166 (9th Cir.

2007) (copyright fair use doctrine “encourages and allows the development of new ideas that build on earlier ones, thus providing a necessary counterbalance to the copyright law’s goal of protecting creators’ work product.”).

Because Anthropic’s use does not show Plaintiffs’ works to end-users, this transformative use of their works in no way deprives Plaintiffs of a market for their books. Instead, Plaintiffs assert a circular theory of harm, that Anthropic has deprived them of the money it would otherwise pay them to license their works in the absence of fair use. But courts reject the notion that failing to license a transformative use weighs against that use being fair. And, there is no evidence that such a market will or even could develop, given the breadth and size of the necessary training corpus, comprising trillions of data points reflecting billions of works.

Nor is there merit to Plaintiffs’ attempt to reduce this case to accusations that Anthropic obtained “pirated works” from “illegal websites.” ECF No. 70 (First Amended Complaint, “FAC”). The Supreme Court has repeatedly warned that such arguments have no place in fair use analysis. What matters is what Anthropic *did* with those books—and making “intermediate” copies to study the relationships among words and concepts in the service of creating a model of how language itself works is a quintessential fair use.

Claude is the type of creative invention that advances the purposes of copyright law. It is transformative in the extreme. The Court should grant Anthropic’s motion for summary judgment on its fair use defense.

SUMMARY OF UNDISPUTED MATERIAL FACTS

A. Anthropic and its mission.

Anthropic is an artificial intelligence (“AI”) company based in San Francisco, California, working to develop generative AI models, and, in particular, LLMs. Declaration of Jared Kaplan (“Kaplan Decl.”) ¶ 6. LLMs are text-based generative AI models trained on extremely large volumes of data to develop a functional understanding of how language works, to use it to generate new text. *Id.* Anthropic’s mission is to build safe, beneficial artificial intelligence. *Id.* ¶ 7.

Since its founding in 2021, Anthropic has extensively published its safety research to foster a race to the top in AI safety. Examples of this research are detailed in the Kaplan Declaration, including

[REDACTED]

Anthropic’s research on interpretability (the science of how LLMs “think”), alignment (the science of developing AI systems that follow human values and intentions), and the societal impacts of LLMs. *Id.* ¶¶ 8-12; www.anthropic.com/research/. During late 2022 through early 2023, Anthropic’s focus evolved from pure research to include commercial deployments of its LLMs, while increasing its rate of publishing safety research. Anthropic focuses on commercial model development alongside safety research because a safe LLM that nobody uses cannot fully demonstrate the potential of reliable, beneficial frontier AI systems. *Id.* ¶¶ 13-15.

B. How people use Claude.

Anthropic’s signature product is a general purpose LLM called Claude. Anthropic has released multiple versions of Claude, beginning with Claude 1, released in March 2023. Its most recent model, Claude 3.7 Sonnet, was released just a few weeks ago. *Id.* ¶¶ 16-17. To create Claude, Anthropic has employed hundreds of engineers who have devoted hundreds of thousands of hours of work to its development, and has spent more than [REDACTED] in computing costs alone related to research and training. Claude assists around [REDACTED] individual users on a daily basis. Anthropic also has [REDACTED] of business customers who rely on Claude. *Id.* ¶¶ 18-19.

Claude is a versatile LLM that can be used in many different contexts depending on a user’s needs. Claude was trained on existing examples of human work product, but it is designed to create new, original outputs and act as a general purpose assistant. *Id.* ¶ 20. Individual users rely on Claude for a vast array of purposes. Anthropic has found that consumers most commonly use Claude to help write computer code and for other business purposes, such as drafting professional emails and analyzing business data, but the diversity of Claude’s uses goes far beyond that. As just a few more examples, businesses use Claude to search deposition transcripts, to evaluate film scripts, to accelerate accounting, to provide multilingual chatbots that expand access to government services, and much more. *Id.* ¶¶ 21-29; <https://www.anthropic.com/customers>.

C. LLM training.

Anthropic’s goal is for Claude to be capable of assisting users with a broad range of capabilities. For this reason, when Anthropic trains Claude, it wants the model to develop a generalized understanding of language patterns and relationships—not to memorize specific

[REDACTED]

expressive content from the data on which it is trained. Kaplan Decl. ¶ 30. Memorization of training data is a problem that inhibits the ability of an LLM to generalize, rather than an intended feature of LLMs. Focusing on the ability to generalize allows LLMs like Claude to understand a virtually limitless array of potential user prompts, and to complete a broad range of tasks by creating original outputs. *Id.* ¶¶ 31-34.

A detailed overview of the process to build and train Claude can be found in the Kaplan Declaration. *See generally id.* ¶¶ 35-68. To briefly summarize, engineers start by building a “neural network,” which is a computational model capable of learning patterns in language and concepts from enormous sets of data, including text called the “training corpus.” *Id.* ¶ 35. These models are called neural networks because they are structured as an interconnected series of nodes (“neurons”) in a layered formation that loosely mimics the structure of the human brain. *Id.* These “nodes” or “neurons” are the model’s basic computational elements, which process information throughout the model, passing bits of words along through the “brain” when the model is asked to do something. *Id.*

The neural network analyzes text in the context of surrounding language in the training data to learn language patterns and the relationship between the words and phrases in the training corpus. *Id.* ¶ 36. Through an iterative process of analyzing the training corpus, the model builds a map of how humans use language and understand concepts without being explicitly programmed with human-defined rules. *Id.* That is, the model itself derives inferences regarding the structure and rules of language from examples provided by engineers. The features of the resulting map are stored within the model in an organized set of numerical values called “parameters,” which effectively track the relationships among words and concepts across the full training corpus—*e.g.*, how strongly a concept like “dog” is associated with related concepts like “puppy,” instead of something unrelated like “smartphone.” *Id.* After training is complete, the model will use its parameters to analyze inputs from users and generate text or perform tasks in response. *Id.* ¶¶ 36-37.

Data Collection and Assembly; Required Volume and Diversity. An LLM like Claude requires a truly massive quantity of data to learn from. Scale is necessary for the statistical, linguistic, and relationship information extracted from the data to be generalizable and for patterns to emerge. *Id.* ¶ 38. The more data used in training LLMs, the better they are able to perform in a wide array of

[REDACTED]

1 subject areas, generalize beyond the corpus, reason, and improve the quality, utility, and creativity of
2 their outputs. *Id.* ¶ 40. Many *trillions* of words are required in order for the model to provide the kinds
3 of flexible, useful, and original responses that Claude can deliver. *Id.* ¶ 39. In Anthropic’s
4 understanding, this scale is effectively a technological requirement, at least given the current state of
5 the art: the company simply could not have built Claude using materially less data than it uses. *Id.*
6 ¶ 41. Put differently, without using the order of magnitude of data that Anthropic used to train Claude,
7 Claude literally could not exist. *Id.*

8 Anthropic measures the volume of training data it uses in “tokens.” This is a term of art
9 referring to the chunks into which Anthropic breaks up text during one of the initial phases of the
10 training process. A token can correspond to a word, a part of a word, or even just particular
11 characters—but all told, on average a given word in the training set will be represented by roughly
12 1.3 tokens. *Id.* ¶ 39. The Claude Sonnet 3.7 model was trained on approximately [REDACTED] “sampled”
13 tokens and [REDACTED] “base” tokens (the number after filtering and deduplication, which does not
14 account for the fact that some tokens are used more than once). This corresponds to roughly [REDACTED]
15 [REDACTED]. *Id.* For reference, this is [REDACTED] the number of tokens available from all the books in the
16 world. Declaration of Steven Peterson (“Peterson Decl.”) ¶ 57.

17 The training data also must be diverse in subject matter and linguistic style for it to train the
18 model effectively. LLMs not trained on diverse training data, including long- and short-form content,
19 will perform poorly in both specific and general ways: they may be able to perform only a limited
20 number of tasks related to that training data, and they will also have a more limited general
21 understanding of facts about the world and about how language works. Kaplan Decl. ¶¶ 42-43.
22 Anthropic’s training corpora are composed of a wide variety of source materials—including scientific
23 papers, computer code, and data from the web—of which books datasets are only a small part, in
24 comparison to much larger volumes of text from across the entire Internet and computer code.
25 *Id.* ¶¶ 45-53.

26 **Pretraining.** After assembling the training corpus, Anthropic takes a number of steps to
27 process the data before showing it to the neural network: deduplicating portions that appear
28 repeatedly, converting the remaining characters to tokens, and randomizing certain sequences for

technical reasons (largely concerning the intention to prevent the model from memorizing portions of the training corpus). *Id.* ¶¶ 54-55. After that, tokens are translated into what are called “vectors,” which are essentially mathematical representations of how words relate to other words. *Id.* ¶¶ 56-59. The model’s weights and biases—numerical values that determine how the model understands vectors and their related concepts to be related—start off as random. *Id.* Through the training process, which involves on the order of a million billion billion repeated mathematical calculations, the model adjusts the weights based on the vector inputs, as it learns more about how humans use words and concepts in writing. *Id.* The representation of that learning is a multidimensional matrix of weights that capture nuances in meaning, such as contextual relationships (how words relate to each other in context—“bank” as in financial institution versus “bank” as in river edge); syntactic information (grammatical properties like parts of speech and word tense); morphological information (information about token structure, prefixes, suffixes, and other word forms); and essentially all the other characteristics of language that are necessary to approximate the way humans use language. This iterative adjustment of the model’s internal weight parameters allows it to learn complex patterns and meaningful concepts regarding language, transforming the initially random vectors into rich, contextual representations of the way language works. *Id.* The outcome of the process is a set of numerical values reflecting the inferences the model has gleaned from the training corpus. The corpus itself is not stored in the model. And once trained, the model no longer uses the corpus. *Id.*

Fine-tuning. A pretrained model is capable of functioning as an LLM, but the quality of the generated outputs from the LLM can be improved—in the sense of making them more useful or relevant to specific tasks—by a process called “fine-tuning.” Fine-tuning is a secondary training phase during which the model is further trained to achieve specific objectives instead of general ones. During fine-tuning, Anthropic implements what it calls “Constitutional AI” principles, which among many other principles of good model behavior, teach Claude to avoid producing outputs that could constitute copyright infringement. *Id.* ¶¶ 60-64.

Inference. Inference is the process of running the LLM. Users supply inputs, like spreadsheets, photos, and/or questions, and Claude generates outputs. Perhaps surprisingly, the same input will not always produce the same output. In the jargon of the industry, that capability makes the

model “probabilistic” rather than “deterministic.” As a result, there is no way to predict exactly how Claude will respond to a given input during the inference process. *Id.* ¶ 65; *see also* ECF No. 79-1 at 33 (Anthropic’s technology tutorial providing examples of probabilistic generation of two haiku about San Francisco’s summer weather).

Guardrails - Additional Suppression of Replication. Finally, Claude also has auxiliary system guardrails designed to stop copyright infringement (among other undesirable behaviors) that sit on top of the model, including its “Prompt Shield” and [REDACTED] *Id.* ¶¶ 66-68.

D. This case is only about use of Plaintiffs’ books as training data.

Plaintiffs are three book authors, who allege that Anthropic infringed their copyrights by using their books in Anthropic’s training corpus for Claude. *See generally* FAC. For purposes of this Motion, Anthropic does not dispute that copies of those books were included in the training corpora for at least one of its commercial Claude models.

The parties agree that this is not a case about whether Claude has produced any output that infringes Plaintiffs’ books. *Id.*; *see also* ECF No. 80 (Technical Tutorial Transcript) at 62:5-7 (Plaintiffs’ counsel stating: “This is not an output copyright infringement [case].”). [REDACTED]

[REDACTED] Ex. 1¹ (Graeber Dep. Tr.) at 24:21-25, 28:5-20; Ex. 2 (Bartz Dep. Tr.) at 34:2-4, 105:20-106:1, 109:2-5; Ex. 3 (Johnson Dep. Tr.) at 35:21-37:21. Plaintiffs also testified that [REDACTED] Ex. 1 (Graeber Dep. Tr.) at 205:11-19; Ex. 2 (Bartz Dep. Tr.) at 218:13-219:6; Ex. 3 (Johnson Dep. Tr.) at 171:11-172:1.

The sole theory of infringement for the Court to address here is whether Anthropic’s use of books to train Claude, in itself, is copyright infringement.

E. Anthropic’s use of books as training data.

Internet Books Datasets. In 2021, when Anthropic first started creating research models, it used certain datasets composed of books that it obtained from the Internet (“Internet books datasets”),

¹ All exhibits are attached to the Declaration of Douglas A. Winthrop in Support of Defendant Anthropic PBC’s Motion for Summary Judgment.

as there were no other sources of books datasets available that were suitable to train LLMs.

Kaplan Decl. ¶¶ 47-50.

Id. ¶ 52; Declaration of Tom Turvey (“Turvey Decl.”)

¶¶ 19-27

Turvey Decl. ¶¶ 22-23, 25.

Id. ¶¶ 26-27.

Kaplan Decl. ¶ 52.

Consistent with the overarching imperative to assemble a very large volume of tokens to train on, when Anthropic includes books datasets in its training corpus, it generally uses the entire text of the books. *Id.* ¶ 47.

F. Licensing is not a practicable approach to assembling the datasets, or even just the books datasets, required to build Claude.

Anthropic also explored whether there was a way to obtain the books-based tokens Claude requires by entering commercial transactions to license electronic books datasets from various potential sources, including from “trade” publishing houses, such as those that publish the Plaintiffs’ books, academic publishers, and self-publishers. It found that there was no viable licensing market that could come close to meeting its needs for books (let alone all the other data required to build a tool like Claude). Turvey Decl. ¶¶ 15-19; 28-49.

² It is also undoubtedly true that some books or portions of books have also been in other places in the training corpora for commercial models, such as Common Crawl, which is a free, open repository of web crawl data. The full Common Crawl dataset includes over 250 billion webpages spanning 18 years. It is essentially a periodic copy of much of the Internet.

Indeed, in some cases, publishers stated that they did not have the rights to license books as LLM training data or were not interested in licensing books for that purpose.

And for the limited licenses that were purportedly on offer, the terms sought by the putative licensors (none of whom owned the rights to the works in suit in this case) were not justifiable given the scale of the data Anthropic needs. *Id.*

Anthropic's experience exploring whether licensing could conceivably work as a way to assemble Claude's training corpus. Each Plaintiff maintains that

Ex. 1 (Graeber Dep. Tr.) at 136:10-22, 222:9-12; Ex. 2 (Bartz Dep. Tr.) at 176:25-177:13; Ex. 3 (Johnson Dep. Tr.) at 123:2-123:12, 145:14-24, 162:15-18.

Not surprisingly then, Exs. 4-6 (RFA No. 9); Ex. 1 (Graeber Dep. Tr.) at 186:11-15; Ex. 2 (Bartz Dep. Tr.) at 208:18-209:7; Ex. 3 (Johnson Dep. Tr.) at 48:24-25.

Ex. 1 (Graeber Dep. Tr.) at 196:3-6; Ex. 2 (Bartz Dep. Tr.) at 208:1-17; Ex. 3 (Johnson Dep. Tr.) at 181:24-183:18.

In fact, Ex. 1 (Graeber Dep. Tr.) at 196:7-24; Ex. 2 (Bartz Dep. Tr.) at 208:18-209:7; Ex. 3 (Johnson Dep. Tr.) at 200:7-14.

Ex. 1 (Graeber Dep. Tr.) at 32:5-37:2, 54:5-23; Ex. 2 (Bartz Dep. Tr.) at 43:3-48:25; Ex. 3 (Johnson Dep. Tr.) at 42:10-58:21.

And even if Anthropic could license books datasets at scale, books only comprise a relatively small part of Anthropic's training corpus for Claude, which is mostly composed of web data from the Internet owned by billions of rightsholders. Kaplan Decl. ¶ 53; Peterson Decl. ¶ 7.


LEGAL STANDARD

The Court should grant Anthropic summary judgment if “there is no genuine dispute as to any material fact and [Anthropic] is entitled to judgment as a matter of law.” Fed. R. Civ. P. 56(a). A dispute is genuine only if there is sufficient evidence for a reasonable fact-finder to find for the non-moving party, and material only if the fact may affect the outcome of the case. *Anderson v. Liberty Lobby, Inc.*, 477 U.S. 242, 247-49 (1986).

ARGUMENT

The Court should grant summary judgment for Anthropic because its copying of Plaintiffs’ books is fair use under the Copyright Act. 17 U.S.C. § 107.

Anthropic copied Plaintiffs’ works solely as a back-end step, invisible to the public, in creating an LLM—a staggeringly complex statistical model of how language works and corresponding facts about the world. Anthropic’s development of this model, plus years of additional product development innovations, yielded a cutting-edge artificial intelligence tool capable of performing work across disciplines and task-types. Plaintiffs make no claim that Claude has produced even a single output that is substantially similar to their books or that Claude is infringing in itself.

A long line of precedent addressing similar back-end copying in developing new digital technologies dictates that such use of copyrighted material is not copyright infringement as a matter of law. To the contrary, “[i]t is precisely this growth in creative expression, based on the dissemination of other creative works and the unprotected ideas contained in those works, that the Copyright Act was intended to promote.” *Sega Enters. Ltd. v. Accolade, Inc.*, 977 F.2d 1510, 1523 (9th Cir. 1992).

To determine whether conduct is fair use and therefore non-infringing, courts consider four non-exclusive factors: “(1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes; (2) the nature of the copyrighted work; (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and (4) the effect of the use upon the potential market for or value of the copyrighted work.” 17 U.S.C. § 107. Applying these factors, courts have concluded that the following uses of copyrighted material were fair uses because the copied works were used only as inputs for making transformative products with uses distinct from the original works:

1 • Scanning all of the millions of books in a dozen or so university libraries to create a searchable
 2 corpus from which users could view short snippets upon request. *Authors Guild v. Google,*
 3 *Inc.*, 804 F.3d 202, 214-25 (2d Cir. 2015) (“*Google Books*”).
 4 • Copying essentially all of the images on the Internet, in order to host thumbnail versions and
 5 display them to users in response to search queries. *Perfect 10, Inc.*, 508 F.3d at 1163-68.
 6 • Duplicating elements of pre-existing computer software in a new tool for developers on a
 7 potentially competing platform. *Google LLC v. Oracle Am., Inc.*, 593 U.S. 1, 26-40 (2021).
 8 The list goes on. *See, e.g., A.V. ex rel. Vanderhye v. iParadigms, LLC*, 562 F.3d 630, 638-45 (4th Cir.
 9 2009) (fair use to copy student papers into a plagiarism detection tool); *Kelly v. Arriba Soft Corp.*,
 10 336 F.3d 811, 817-22 (9th Cir. 2003) (fair use to copy the entire universe of images on the Internet
 11 for search purposes); *Sega*, 977 F.2d at 1522-27 (fair use to copy proprietary operating system in
 12 order to create competing video games); *Sony Comput. Ent., Inc. v. Connectix Corp.*, 203 F.3d 596,
 13 601-08 (9th Cir. 2000) (fair use to “repeatedly cop[y]” proprietary operating system to create
 14 unauthorized platform for PlayStation games). What all these cases have in common is that, like
 15 Anthropic’s use, they involved use of copyrighted works only as inputs in making new, non-
 16 infringing, and transformative products that served different purposes than the original works.

17 Applying established precedent to the undisputed facts here yields the same result. Copyright
 18 law does not give Plaintiffs the right to prevent Anthropic from making copies in order to study
 19 Plaintiffs’ writing, extract uncopyrightable information from it, and use what it learned to create
 20 revolutionary technology that itself does entirely new things. Doubly so where, as here, no market
 21 does or could exist to meet the technological requirements of the new product via licensing
 22 transactions.

23 **I. THE FIRST FACTOR FAVORS FAIR USE BECAUSE USING COPYRIGHTED**
 24 **WORKS TO TRAIN AN LLM IS QUINTESSENTIALLY TRANSFORMATIVE**

25 The first fair use factor—the purpose and character of the new use—“lies at the heart of the
 26 fair user’s case.” Pierre N. Leval, *Toward A Fair Use Standard*, 103 Harv. L. Rev. 1105, 1111 (1990)
 27 (“Leval on Fair Use”); *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569 (1994) (favorably citing
 28 Judge Leval’s article extensively). The inquiry distinguishes between uses that substantially substitute

[REDACTED]

1 for original works, and those that serve new or further end-uses that are often labeled
2 “transformative.” *See Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, 598 U.S. 508, 527-
3 29 (2023). Transformative uses either (i) create “something new, with a further purpose or different
4 character,” or (ii) “expand” the original’s utility to further copyright’s objective of contributing to
5 public knowledge. *Google Books*, 804 F.3d at 214, 219 (citations omitted).

6 Anthropic’s use of Plaintiffs’ works was transformative in each of the several senses the cases
7 have recognized. Assembling the datasets, tokenizing it, studying the interrelationships among the
8 tokens, and using the information gleaned from those interrelationships to create an LLM exemplifies
9 literal transformation. The model itself is not a repository of copies of the works in the training set.
10 Kaplan Decl. ¶¶ 56, 59. Instead, the process breaks the words apart and analyzes them, with
11 information from and about the training data then represented in the resulting model only as numerical
12 parameters reflecting what the model has learned from the training corpus. *Id.* ¶¶ 35-37, 56-57. This
13 transformation creates an infinitely extensible tool that, rather than being limited by the particular
14 expression of ideas contained in its training corpus, responds to novel questions with original answers.
15 *Id.* ¶¶ 20-21. An LLM differs fundamentally from a book. It is a generative software tool, not a static
16 text. This difference alone proves transformativeness under the first fair use factor. Peterson Decl. ¶¶
17 42-44.

18 But Anthropic’s new use is also transformative because it serves a purpose different from the
19 pre-existing works’ original one, adding new utility. *See Oracle*, 593 U.S. at 30 (noting in first-factor
20 analysis that the defendant used the plaintiff’s work “to create new products” by providing a “highly
21 creative and innovative tool”). As the Second Circuit explained in *Google Books* in affirming that it
22 was fair use to scan millions of books to enable many otherwise impossible new functionalities, the
23 first fair use factor strongly favors *enabling*, not thwarting, innovation that extracts uncopyrightable
24 information from and about copyrighted works. 804 F.3d at 217. Similarly, *Authors Guild, Inc. v.*
25 *HathiTrust* held that creating a searchable database of millions of books was “quintessentially
26 transformative” because it enabled a new kind of functionality with entirely different “purpose,
27 character, expression, meaning, and message” from the original works. 755 F.3d 87, 97 (2d Cir. 2014).

Anthropic’s new use is meaningfully more transformative than those examples. *Google Books* found transformativeness in technology that showed the “frequency of usage of selected words in the aggregate corpus of published books,” (804 F.3d at 217), and *HathiTrust* focused on the utility of searchable databases, (755 F.3d at 97-98). Claude goes further: it does not merely report word frequency data or simply allow users to search an assembled corpus of works at all. Instead, Claude uses data resulting from extensive analysis of that corpus to create a model of language itself. That makes Claude itself transformational. And the transformative purpose and use does not even end there: Claude assists humans across myriad fields, from software development to medical science, enabling its end users to drive innovation in countless new ways. It is undisputed that Plaintiffs’ books cannot write code. It is undisputed that Plaintiffs’ books cannot synthesize data and draft clinical trial reports. It is undisputed that Plaintiffs’ books cannot engage in human-like conversation and iteration. Claude is clearly something new. Kaplan Decl. ¶¶ 20-21.

Warhol, the Supreme Court’s most-recent fair use decision, reinforces the transformative nature of Anthropic’s model training. The case held that the first factor weighed against fair use where the use in question served essentially the same purpose as that for which the original work was created. Specifically, licensing Andy Warhol’s depiction of the musician Prince for publication in a magazine was a use with a purpose indistinguishable from the purpose of the source photograph. *See Warhol*, 598 U.S. at 535-38 (explaining that both works were “portraits of Prince used in magazines to illustrate stories about Prince”). In that circumstance, the Court explained, the copy directly substituted for use of the original in a long-established market. *Id.* at 537-38. But in so ruling, the Court made clear what *would* be transformative: the use of a work serving a “distinct purpose” than that for which it was created, without supplanting demand for the originally intended use. *Id.* at 510-11. Such transformative use would be “justified because it furthers the goal of copyright, namely, to promote the progress of science and the arts, without diminishing the incentive to create.” *Id.* In the dichotomy the Court described, Anthropic’s use plainly falls on the “transformative” side.

Decades of precedent adjudicating fair use in the context of intermediate copying applies the same foundational principles in the same way. Even beyond the directly apposite book-scanning and image-search decisions, plaintiffs in copyright cases commonly contend that it is infringement for

defendants to copy plaintiffs' works for the purpose of developing some new technological tool. Courts routinely reject those claims, notwithstanding the defendants' back-end use of the copyrighted works, when the resulting user-facing technology is itself non-infringing. That is what happened in *Sega*, in which the defendant copied the plaintiff's operating system in order to figure out how to create unauthorized but non-infringing video games that could be played on the plaintiff's gaming console. 977 F.2d at 1522-27. It is what happened in *Sony*, in which the defendant engaged in similar copying to create its own emulator on which authorized games for the proprietary platform could be played. 203 F.3d at 606-07. And it is what happened in *iParadigms*, in which the defendant used the plaintiff's copyrighted school work to develop a plagiarism-detection tool. 562 F.3d at 638-40.

Thomson Reuters Enter. Ctr. GMBH v. ROSS Intelligence Inc., 2025 WL 458520 (D. Del. Feb. 11, 2025), does not require a different result. That decision—an unexpected reversal of that court's own prior ruling (694 F. Supp. 3d 467, 486 (D. Del. 2023))—held that the defendant's copying of Westlaw headnotes was not transformative because the defendant used the plaintiff's works for precisely the same purpose as the plaintiff—to create a legal research tool—to directly compete with the plaintiff. 2025 WL 458520, at *7-8. The court also emphasized that the defendant's product was not generative AI, but rather a tool that “spits back relevant judicial opinions” similar to Westlaw headnotes. *Id.* at *7. Here, in contrast, Claude serves very different purposes from books, does not compete with them, and produces transformative, generative content. *Accord Concord Music Grp., Inc., et al. v. Anthropic PBC*, No. 24-cv-03811-EKL (N.D. Cal. March 25, 2025) (ECF 321), at 1 n.1 (declining to consider *ROSS* in denying music publishers' motion for preliminary injunction because, *inter alia*, “it did not concern a generative AI model; and . . . the parties in that case were direct competitors”).

It is no accident that the overwhelming number of cases in this area come out the way they do: the point of copyright law is to incentivize innovation, and the law accordingly favors uses of copyrighted works that further that goal. Courts have consistently applied this basic principle about copyright's primary purpose—to encourage and make space for technological progress and further creative expression—specifically to conclude that the first factor favors fair use, in cases involving new technologies built in part using third parties' copyrighted works. The Supreme Court itself did

so in *Oracle*. See 593 U.S. at 30 (“To the extent that Google used parts of the Sun Java API to create a new platform that could be readily used by programmers, its use was consistent with that creative ‘progress’ that is the basic constitutional objective of copyright itself.”). And that holding built on a rich tradition of appellate court cases taking the same approach. See, e.g., *Kelly*, 336 F.3d at 820 (“[T]his first factor weighs in favor of Arriba due to the public benefit of the search engine and the minimal loss of integrity to Kelly’s images.”); *Perfect 10*, 508 F.3d at 1166-67 (relying on legislative history to emphasize the flexible nature of fair use “especially during a period of rapid technological change,” in holding that the first factor strongly favored fair use); *Sega*, 977 F.2d at 1523 (recognizing the first factor favors technologies enabling a “growth in creative expression . . . that the Copyright Act was intended to promote”).

Finally, neither commerciality nor any purported bad faith undermines Anthropic’s fair use defense. As to commerciality, Anthropic has used Plaintiffs’ works to train both noncommercial research and development models and commercial models. But even as to Anthropic’s uses of Plaintiffs’ works for commercial purposes, courts consistently hold that high degrees of transformativeness trump defendants’ commercial purposes. One example is *Oracle*, which found that even though Google’s use was commercial, that fact was not dispositive because its technology was “inherently transformative.” *Oracle*, 593 U.S. at 32. And a litany of cases supports the principle that *Oracle* applied—that transformative uses are still fair even if they are commercial. See *Campbell*, 510 U.S. at 573-74, 584-85 (reversing appellate court for putting too much weight on commerciality of use); *Google Books*, 804 F.3d at 219 (noting that the Second Circuit has “repeatedly rejected the contention that commercial motivation should outweigh a convincing transformative purpose”); *Apple Inc. v. Corellium, Inc.*, 2023 WL 3295671, at *9 (11th Cir. May 8, 2023) (“[M]any fair uses are commercial.”).

As for “bad faith,” it is unclear whether Plaintiffs contend that a defendant’s purported bad faith in acquiring a work precludes fair use altogether, that it weighs against fair use, or something else. So far, it appears that Plaintiffs plan to argue that acquiring training data from so-called “shadow libraries” precludes a fair use defense. The law does not support any version of that argument. The Supreme Court has now *twice* expressed skepticism that “bad faith has *any* role in a fair use analysis.”

[REDACTED]

1 *Oracle*, 593 U.S. at 32 (emphasis added) (citing *Campbell*, 510 U.S. at 585 n.18; *Leval* on Fair Use
2 1126). That echoes this Court’s conclusion a decade ago that a “respectable view” exists that good or
3 bad faith should no longer be a consideration after *Campbell*. *Oracle Am., Inc. v. Google Inc.*, 2016
4 WL 3181206, at *2 (N.D. Cal. June 8, 2016) (Alsup, J.) (“[E]ither a use is objectively fair or it is not
5 and subjective worry over the issue arguably should not penalize the user.”). We submit that the
6 “respectable view” expressed by this Court and endorsed by the Supreme Court in *Oracle* is the law.

7 But even assuming bad faith plays some role in the operative analysis, it could not be a
8 material factor here. The Ninth Circuit has held that acquiring unauthorized images posted by third
9 parties on the Internet without the copyright owners’ permission does not preclude a finding of fair
10 use. *Perfect 10*, 508 F.3d at 1164 n.8. In fact, in *Time Inc. v. Bernard Geis Associates*, on which the
11 Supreme Court relied in *Harper & Row Publishers, Inc. v. Nation Enterprises*, 471 U.S. 539, 562-63
12 (1985), the court granted summary judgment on fair use notwithstanding the defendant’s bad faith
13 conduct of sneaking into the plaintiff’s office and copying the plaintiff’s film. 293 F. Supp. 130, 136,
14 146 (S.D.N.Y. 1968). These cases show that if bad faith has any weight, it is minimal.

15 [REDACTED]
16 [REDACTED] Turvey Decl.
17 ¶¶ 20-27.
18 [REDACTED] See Ex. 1 (Graeber
19 Dep. Tr.) at 102:21-103:9; Ex. 2 (Bartz Dep. Tr.) at 125:2-7; *see also* Ex. 3 (Johnson Dep. Tr.) at
20 41:19-21; *accord Time Inc.*, 293 F. Supp. at 146 (bad faith acquisition did not defeat fair use when
21 defendant “could have” acquired work by other means). Even putting the maximum weight possible
22 on Plaintiffs’ allegations regarding so-called “shadow libraries,” as a matter of law, those facts cannot
23 outweigh the highly transformative nature of Anthropic’s use of Plaintiffs’ works.

24 * * *

25 Anthropic’s use of copyrighted content to train its LLMs is a paradigmatic transformative use,
26 and the first factor accordingly strongly favors fair use.

II. THE SECOND FACTOR FAVORS FAIR USE BECAUSE ANTHROPIC USES PLAINTIFFS' PUBLISHED WORKS FOR A TRANSFORMATIVE PURPOSE

The second fair use factor looks to “the nature of the copyrighted work.” 17 U.S.C. § 107(2). This factor “typically has not been terribly significant in the overall fair use balancing.” *Mattel, Inc. v. Walking Mountain Prods.*, 353 F.3d 792, 803 (9th Cir. 2003) (citation omitted). While this factor is generally protective of “creative works” close to copyright’s core, (*Kelly*, 336 F.3d at 820), it nevertheless tilts towards fair use when a “creative work . . . is being used for a transformative purpose,” (*HathiTrust*, 755 F.3d at 98 (alteration in original) (citation omitted)). Moreover, whether a work is “unpublished” or not is also a “critical element of its nature.” *Kelly*, 336 F.3d at 820. Uses of published works “are more likely to qualify as fair use because the first appearance of the artist’s expression has already occurred.” *Id.*

Here, although Plaintiffs’ works are creative, they are also published and indisputably widely available. Because Anthropic uses these works for transformative purposes “rather than replicating protected expression in a manner that provides a meaningful substitute for the original[s],” (*Google Books*, 804 F.3d at 220), the second factor also favors fair use.

III. THE THIRD FACTOR FAVORS FAIR USE BECAUSE ANTHROPIC’S TRANSFORMATIVE PURPOSE REQUIRES COPYING ENTIRE WORKS

The third fair use factor looks to the “amount and substantiality of the portion used” of the copyrighted work. 17 U.S.C. § 107(3). “For some purposes, it may be necessary to copy the entire copyrighted work.” *HathiTrust*, 755 F.3d at 98 (citing *Bill Graham Archives v. Dorling Kindersley Ltd.*, 448 F.3d 605, 613 (2d Cir. 2006)). Thus, “[e]ntire verbatim reproductions are justifiable where the purpose of the work differs from the original.” *Mattel*, 353 F.3d at 803 n.8. The third factor will favor fair use unless the “copies are excessive or unreasonable in relation to the purposes identified” by the defendant. *HathiTrust*, 755 F.3d at 99; see *Kelly*, 336 F.3d at 821 (finding wholesale copying “reasonable”); *Google Books*, 804 F.3d at 221-22 (complete copying “justified as fair use when the copying was reasonably appropriate to achieve the copier’s transformative purpose”).

In permitting wholesale copying, courts have found it fair to copy libraries full of books (*Google Books*, 804 F.3d at 221; *HathiTrust*, 755 F.3d at 98), entire briefs to make them text

searchable (*White v. W. Pub. Corp.*, 29 F. Supp. 3d 396, 399 (S.D.N.Y. 2014)), and entire student papers to create a tool to detect plagiarism (*iParadigms*, 562 F.3d at 634). And the Ninth Circuit held in *Sega*, 977 F.2d at 1522-23, and *Sony*, 203 F.3d at 605-06, that where the copying was only intermediate, the fact that the defendants made an internal copy of the entire work in order to extract what they needed for their transformative purpose weighed little against fair use.

Similarly, Anthropic cannot achieve its transformative purpose by copying only portions of books. For Anthropic's LLMs to function effectively, Anthropic's models require trillions of tokens, equivalent to the amount of text in [REDACTED] Kaplan Decl. ¶ 39; Peterson Decl. ¶¶ 15, 56. Moreover, to be effective the model requires context from diverse sources, including longer-length documents. Kaplan Decl. ¶¶ 42-43. Anthropic's copying is therefore not excessive or unreasonable because it is necessary to achieve its transformative purpose. The third factor favors fair use.

IV. THE FOURTH FACTOR FAVORS FAIR USE BECAUSE ANTHROPIC'S TRANSFORMATIVE COPYING DOES NOT HARM ANY COGNIZABLE MARKET

The fourth factor considers "the effect of the use upon the potential market for or value of the copyrighted work." 17 U.S.C. § 107(4). In this case, it overwhelmingly favors fair use: first, as a matter of law, there is no right to extract licensing fees for transformative uses of copyrighted works. Second, on the facts, there is no viable theory of market harm. There is no harm to any existing, traditional market by which Plaintiffs exploit their books, *e.g.*, by selling copies to the public. And with respect to the contention that harm arises from the failure to license copyrighted works as training data for LLMs, no such market exists or plausibly could, because several different forms of market failure preclude it. Third, as in *Oracle*, the public benefit of the use in question outweighs any theoretical harm to the value of the copyrighted works. *See* 593 U.S. at 35-38.

A. **There is no cognizable market harm from Anthropic's transformative use as a matter of law.**

Courts have long recognized that certain types of harm from copying are not "cognizable under the Copyright Act." *Oracle*, 593 U.S. at 35 (quoting *Campbell*, 510 U.S. at 592). These authorities hold that "any economic 'harm' caused by transformative uses" does not "count" for purposes of the fourth factor analysis because transformative uses "do not serve as substitutes for the

original work,” (*HathiTrust*, 755 F.3d at 99), and because “[c]opyright owners may not preempt exploitation of transformative markets,” (*Bill Graham Archives*, 448 F.3d at 615 (citation omitted)).

Thus, “a copyright holder cannot prevent others from entering fair use markets” by claiming harm associated with lost licensing fees from a use that is itself transformative. *Tresóna Multimedia, LLC v. Burbank High Sch. Vocal Music Ass’n*, 953 F.3d 638, 652 (9th Cir. 2020) (citation omitted).

Following these precedents, Plaintiffs’ fourth-factor theory fails as a matter of law. Their principal contention seems to be that the market harm they have suffered is harm from Anthropic’s failure to pay them licensing fees for using their works as training data—not that the use of their books results in traditional book-buyers eschewing purchases in favor of accessing the same content from Claude. But that contention falls prey to a long-recognized circularity problem: “[b]y definition every fair use involves some loss of royalty revenue because the secondary user has not paid royalties.” *Bill Graham Archives*, 448 F.3d at 615 (quoting Leval on Fair Use 1124). Courts squarely reject attempts to bootstrap factor-four market harm from the mere fact that the defendant did not pay the plaintiff for the very use in question when the use is transformative.

Rather than looking to such circular theories of market harm, under the fourth factor “[t]he only market harms that count are the ones that are caused because the secondary use serves as a substitute for the original, not when the secondary use is transformative.” *HathiTrust*, 755 F.3d at 99; *see also Seltzer v. Green Day, Inc.*, 725 F.3d 1170, 1179 (9th Cir. 2013); *Google Books*, 804 F.3d at 223-25; *Campbell*, 510 U.S. at 591 (recognizing a clear connection between the first and fourth fair use factors). But there is no evidence that Claude substitutes for Plaintiffs’ books or that anyone is using Claude in lieu of reading Plaintiffs’ books. Peterson Decl. ¶¶ 21-22.

Plaintiffs have argued that Anthropic’s use of their books as training data does ultimately substitute for end-user consumption of their books in that someone could use Claude to write books that compete with theirs. *See* FAC ¶¶ 52-53. That contention also fails. There is no record evidence showing that books generated by Claude, if any exist, compete with Plaintiffs’ books. Peterson Decl. ¶¶ 31-33. And even if there were any such evidence—that is, evidence that someone used Claude to write a book that resulted in lost sales of Plaintiffs’ books—it would not be the kind of substitution that is cognizable under the fourth factor.

[REDACTED]

1 *Sega* addressed this very issue. 977 F.2d at 1523. There, the defendant made copies of the
2 plaintiff's work specifically to enable creation of video games that would compete with the plaintiff's
3 pre-existing video games. The Ninth Circuit nevertheless found that the fourth factor favored fair use,
4 both because it made no sense to assume that someone who bought one video game would not buy
5 two, and because impeding the creation of new works "runs counter to the statutory purpose of
6 promoting creative expression and cannot constitute a strong equitable basis for resisting the
7 invocation of the fair use doctrine." *Id.* at 1523-24 ("Mike Ditka Power Football" video game did not
8 substitute for "Joe Montana Football" video game); *see Sony*, 203 F.3d at 607-08 (fair use favored
9 despite economic loss to plaintiff). Because Anthropic's use of Plaintiffs' works is transformative,
10 furthers the statutory purpose of promoting new creative expression, and "promote[s] the Progress of
11 Science and the useful Arts," its use is fair regardless of any theoretical economic impact on any
12 market for Plaintiffs' works. *Oracle*, 593 U.S. at 30 (quoting U.S. Const. art. 1, § 8, cl. 8).

13 **B. There is no viable existing or potential market to license Plaintiffs' works for**
14 **LLM training.**

15 Even if the law permitted Plaintiffs to bootstrap a fourth-factor market harm theory from the
16 failure to license training data by itself, no such market exists or plausibly could.

17 Plaintiffs admit that none of their works has ever been licensed to train LLMs. Exs. 4-6 (RFA
18 No. 9). In fact, [REDACTED]

19 [REDACTED]
20 [REDACTED] Ex. 1 (Graeber Dep. Tr.) at 196:7-24; Ex. 2 (Bartz Dep. Tr.) at 208:18-209:7; Ex. 3 (Johnson
21 Dep. Tr.) at 200:7-14. "[I]f a copyright holder has not made a license available to use a particular
22 work in a particular manner, the inference is that the author or publisher did not think that there would
23 be enough such use to bother making a license available." *Cambridge Univ. Press v. Patton*, 769 F.3d
24 1232, 1277 (11th Cir. 2014); *see also Bell v. Eagle Mountain Saginaw Indep. Sch. Dist.*, 27 F.4th 313,
25 325-26 (5th Cir. 2022) (finding fair use when plaintiff alleged no market for licensing his work other
26 than filing lawsuits and entering into settlement agreements). Accordingly, there is no existing market
27 for Plaintiffs' works for the use in question.

28 Nor is a market for training data potentially "reasonable, or likely to be developed." *Seltzer*,

[REDACTED]

725 F.3d at 1179. That issue turns on a range of considerations, including whether there are purely practical impediments to a market emerging. Intuitively, fair use is favored when “the copyright owner [cannot] command a license fee commensurate with the costs of transacting with the copier.” *Ty, Inc. v. Publ’ns Int’l Ltd.*, 292 F.3d 512, 518 (7th Cir. 2002); *see also Bell v. Milwaukee Bd. of Sch. Dirs.*, 2022 WL 18276966, at *7 (E.D. Wis. Dec. 21, 2022) (fourth factor favors fair use where “the transaction costs associated with . . . licensing would far outweigh the value of the” use).

This is not the first case to address whether a potential licensing market is harmed when the plaintiff’s works were a small fraction of a vast corpus. In *White v. West Pub. Corp.*, for example, a lawyer filed a putative class action against West and Lexis for including briefs he had written in their electronic databases. 29 F. Supp. 3d 396, 397 (S.D.N.Y. 2014). The databases contained twelve million works (*id.* at 398), a fraction of the data used to train Claude. The court ruled “that no potential market exists because the transaction costs in licensing attorney works would be prohibitively high.” *Id.* at 400.

Licensing Claude’s training corpus, or even just the books in Claude’s training corpus, would present comparatively greater challenges, by orders of magnitude. Consider first the Plaintiffs’ own testimony [REDACTED]

[REDACTED] Ex. 1 (Graeber Dep. Tr.) at 136:10-16, 137:13-15; Ex. 2 (Bartz Dep. Tr.) at 176:25-177:13; Ex. 3 (Johnson Dep. Tr.) at 122:18-25, 145:14-20, 161:3-7.

And all three Plaintiffs testified that [REDACTED]

[REDACTED] Ex. 1 (Graeber Dep. Tr.) at 33:1-20; Ex. 2 (Bartz Dep. Tr.) at 43:9-17; Ex. 3 (Johnson Dep. Tr.) at 42:6-19.

This is not the foundation for a functional future licensing market for the use in question—irrespective of whether the use in question is framed as licensing Plaintiffs’ particular works, licensing books as a category for use as training data for an LLM like Claude, or licensing the full corpus of data required to develop Claude. On average, a book provides 100,000 or so of the [REDACTED] words needed to train Claude (about [REDACTED]). The incremental value of licensing a single book (or a single author’s books) for training would be dwarfed by the costs of negotiation.

[REDACTED]

Peterson Decl. ¶¶ 7, 65-68. Replicating that experience to negotiating the volume of licenses that would be necessary would be even more impractical. Even just identifying and getting contact information for millions of book authors or owners to secure rights to the trillions of tokens Anthropic requires in books, by itself, would be an utterly impossible task, to say nothing of engaging in and concluding millions of contract negotiations. Peterson Decl. ¶¶ 62, 66-67. Plus, there is the time involved: even if Anthropic could consummate book licenses at the wildly optimistic rate of 1,000 per day, working seven days per week, it still would take more than 2.7 years to sign one million licenses. Assuming 100,000 tokens per license, that number of licenses meets only [REDACTED] of Anthropic's need for at least [REDACTED] tokens. Peterson Decl. ¶ 56.

The purely logistical obstacles to licensing training data at scale are not merely hypothetical. While licensing from individual authors is obviously impractical, Anthropic broadly explored whether it would be possible to compile the books required for Claude's development by licensing through publishers. But that approach was unworkable. A number of the publishers did not have the rights required to conclude licensing deals. Notably, none of Plaintiffs' publishers—Simon & Schuster, Hachette, and Penguin Random House—offered Anthropic any licenses at all. In total, there were not nearly enough books licensable through publishers to generate the volume of books data required by Anthropic. The few publishers who could in fact license their content offered relatively small volumes of data on terms that were not justifiable for Anthropic based on the scale of its needs. Turvey Decl. ¶¶ 15-19; 28-49.

The operative complaint in this case cites a handful of licenses that other AI companies have allegedly entered into. FAC ¶ 48. Even if there were evidence to support the complaint's allegations, that does not prove the existence of a workable potential market for licensing LLM training data at scale. Whatever the motivation for those deals—whether averting threatened litigation or otherwise—their existence does not solve the practical barriers to being able to build LLMs without using unlicensed data. The cited agreements evidently cover only a tiny fraction of the training data that AI companies use, requiring those AI firms themselves to continue using unlicensed training data and to face the resulting onslaught of copyright litigation. As a result, those deals—to the extent they are what Plaintiffs say they are—confirm the futility of attempting to license the amount of tokens

[REDACTED]

1 necessary to train an LLM comparable to Claude. Peterson Decl. ¶¶ 56-59.

2 The fact that these licensing deals only provide a small amount of data is significant because
3 a licensing market is “likely to be developed” only if the market provides Anthropic a sufficient
4 number of tokens to train its LLMs. *See Seltzer*, 725 F.3d at 1179. The relevant inquiry of whether
5 transaction costs can prevent a market from forming is not limited to the transaction costs of
6 negotiating with just Plaintiffs, or even just all owners of book copyrights, but the transaction costs
7 of licensing all the diverse data necessary to meet Claude’s need for [REDACTED] of tokens. *See*
8 *White*, 29 F. Supp. 3d at 398, 400.

9 Finally, Anthropic’s needs for [REDACTED] of tokens cannot be met by books alone, for
10 reasons of both scale and diversity. That is why, looking at the other sources of Anthropic’s data only
11 further confirms any licensing market will fail. For example, Anthropic has used Common Crawl, a
12 dataset containing over 250 billion webpages that is essentially a periodic copy of much of the Internet
13 over the past 18 years. Tracking down every owner of every copyrighted work on the Internet and
14 negotiating licenses with billions of such content owners is not remotely possible. Thus, even if it
15 were somehow practicable to negotiate licenses for all the world’s copyrighted books, no licensing
16 market would develop because the market would not provide enough tokens to train Anthropic’s
17 LLMs, eliminating Anthropic’s incentive to purchase any licenses at all. Peterson Decl. ¶¶ 56-59.

18 In sum, the practical impediments and transaction costs to develop a viable market that could
19 support the use in question are insurmountable, and the fourth factor favors fair use on this basis as
20 well.

21 **C. The public benefits of Claude outweigh any potential market effect.**

22 Even if Plaintiffs could establish harm to a cognizable licensing market, the fourth factor still
23 favors fair use because “the public benefits the copying will likely produce” outweigh any such
24 theoretical harm. *Oracle*, 593 U.S. at 35; *see Perfect 10*, 508 F.3d at 1166 (weighing plaintiff’s alleged
25 harm against “the interests of the public”). Courts consider the public benefits of copying to prevent
26 existing rightsholders from “stamp[ing] out the very creativity that the Act seeks to ignite.” *SOFA*
27 *Ent., Inc. v. Dodger Prods., Inc.*, 709 F.3d 1273, 1278 (9th Cir. 2013).

28 Here, the public receives vast benefits from Claude, as people around the world use it for

inventive and important purposes. For example:

- Norvo Nordisk uses Claude to reduce the time spent producing clinical report data from 12 weeks to 10 minutes, speeding up the approval process for new treatments.
- The University of Sydney uses Claude to accelerate whale conservation. Claude analyzes immense amounts of acoustic data, which conservationists use to learn about the location and migration of minke whales.
- The European Parliament used Claude to improve access to 2.1 million archived government records dating back to 1952. Claude can search, summarize, translate, build reports, and extrapolate information about the documents.
- Magic School uses Claude to generate curriculum plans, quizzes, and handle administrative tasks for teachers so that they can focus on teaching students.

Kaplan Decl. ¶¶ 25-28. Claude has many other uses benefiting people from a wide array of professions and walks of life, including searching deposition transcripts, analyzing Hollywood film scripts, accelerating accounting tasks, and powering a chatbot that allows city employees and residents to access information in English and Spanish about health services. *Id.* ¶ 29.

The immense benefits that Claude provides to the public have not caused a single person to forgo buying any one of Plaintiffs' books. Those benefits do not flow from selling the public copies of Plaintiffs' copyrighted expression, but rather from Claude's ability to transform a mass of data into new expression and enable a dizzying number of useful purposes. In balancing Plaintiffs' legitimate interest under the Copyright Act in preventing others from supplanting their works against the immense benefits of Claude—which does not supplant Plaintiffs' works at all—the balance is overwhelmingly in favor of fair use. Finding fair use here would foster innovation, expand access to knowledge, and create powerful new tools for human creativity—precisely the outcomes copyright law was designed to encourage.

CONCLUSION

For the foregoing reasons, the Court should grant Anthropic's motion and enter judgment for Anthropic.

1 Dated: March 27, 2025

Respectfully submitted,

2 **ARNOLD & PORTER KAYE SCHOLER LLP**

3
4 By: /s/ Douglas A. Winthrop

5 DOUGLAS A. WINTHROP

6 *Attorneys for Defendant*

7 ANTHROPIC PBC



CERTIFICATE OF SERVICE

I, Douglas A. Winthrop, am the ECF user whose identification and password are being used to file the foregoing **DEFENDANT ANTHROPIC PBC'S NOTICE OF MOTION AND MOTION FOR SUMMARY JUDGMENT; MEMORANDUM OF POINTS AND AUTHORITIES IN SUPPORT; AND SUPPORTING DECLARATIONS AND EXHIBITS.**

Dated: March 27, 2025

/s/ Douglas A. Winthrop